AEM
Journals.ASM.org

Annotation of the Modular Polyketide Synthase and Nonribosomal Peptide Synthetase Gene Clusters in the Genome of *Streptomyces tsukubaensis* NRRL18488

Marko Blažič,[b] Antonio Starcevic,[a] Mohamed Lisfi,[d] Damir Baranasic,[a] Dušan Goranovič,[c] Štefan Fujs,[c] Enej Kuščer,[c,e] Gregor Kosec,[c,e] Hrvoje Petković,[c]* John Cullum,[d] Daslav Hranueli,[a] and Jurica Zucko[a]

Faculty of Food Technology and Biotechnology, University of Zagreb, Zagreb, Croatia[a]; University of Ljubljana, Biotechnical Faculty, Department of Food Science and Technology, Ljubljana, Slovenia[b]; Acies Bio d.o.o., Ljubljana, Slovenia[c]; Department of Genetics, University of Kaiserslautern, Kaiserslautern, Germany[d]; and Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins (CIPKeBiP), Ljubljana, Slovenia[e]

The high G+C content and large genome size make the sequencing and assembly of *Streptomyces* genomes more difficult than for other bacteria. Many pharmaceutically important natural products are synthesized by modular polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs). The analysis of such gene clusters is difficult if the genome sequence is not of the highest quality, because clusters can be distributed over several contigs, and sequencing errors can introduce apparent frameshifts into the large PKS and NRPS proteins. An additional problem is that the modular nature of the clusters results in the presence of imperfect repeats, which may cause assembly errors. The genome sequence of *Streptomyces tsukubaensis* NRRL18488 was scanned for potential PKS and NRPS modular clusters. A phylogenetic approach was used to identify multiple contigs belonging to the same cluster. Four PKS clusters and six NRPS clusters were identified. Contigs containing cluster sequences were analyzed in detail by using the ClustScan program, which suggested the order and orientation of the contigs. The sequencing of the appropriate PCR products confirmed the ordering and allowed the correction of apparent frameshifts resulting from sequencing errors. The product chemistry of such correctly assembled clusters could also be predicted. The analysis of one PKS cluster showed that it should produce a bafilomycin-like compound, and reverse transcription (RT)-PCR was used to show that the cluster was transcribed.

Bacteria of the genus *Streptomyces* are characterized by their complex secondary metabolism and are known to be prolific producers of many clinically useful compounds with antibacterial, antifungal, anticancer, immunosuppressive, and other activities. The biosynthesis of many of these bioactive compounds is encoded in the genomes of *Streptomyces* strains in the form of biosynthetic clusters containing relatively large genes with a modular structure, where each module is involved in the catalysis of a particular biosynthetic step. Modular polyketide synthases (PKSs), for example, produce erythromycin (antibiotic), avermectin (antiparasitic), and spinosyn (insecticide). Modular nonribosomal peptide synthetases (NRPSs) produce ACV tripeptide [δ-(L-α-amino-adipate)-L-cysteine-D-valine; precursor of penicillins and cephalosporins] and vancomycin (antibiotics) as well as numerous larger compounds, such as cyclosporine (immunosuppressant). Furthermore, hybrid PKS-NRPS modular enzymes mediate the biosynthesis of epothilone (cytostatic) and several immunosuppressants, such as rapamycin and FK506 (tacrolimus), which are structurally related. Biosynthetic enzymes which catalyze the production of these compounds consist of a series of modules usually distributed among several polypeptides. Each module consists of a series of enzymatic domains involved in the catalysis of individual biosynthetic steps. During biosynthesis, the nascent polyketide or peptide chain is passed from module to module, and each module adds an extender unit until synthesis is complete. The modular nature of the synthesis makes it possible to make predictions about the chemical nature of the product from the cluster DNA sequence (26).

Generally, natural products have been discovered and characterized through a multistep process, usually initiated by the screening of fermentation broths of natural microbial isolates for desired activities. This was followed by the purification of chemical entities and their structure determination. The majority of clinically useful compounds were discovered in this way. A key step in this process is the development of industrial strains with an increased production of the target compounds by using the laborious methods of classical mutagenesis and selection (1). This approach is very time-consuming and has two major disadvantages: production levels are very dependent on the fermentation conditions, so many metabolites will be missed, and known natural products are often reisolated. The majority of the *Streptomyces* genomes contain several distinct modular biosynthetic clusters so that each strain is capable of biosynthesizing a range of chemically diverse natural products (3, 13, 22). While multiple clusters may be expressed simultaneously, often resulting in impurities which hamper the efficiency of industrial bioprocesses for desired compounds, the majority are transcribed at low levels if at all. Such clusters are often referred to as "silent" or "cryptic" clusters, and

Received 15 June 2012 **Accepted** 9 September 2012
Published ahead of print 14 September 2012
Address correspondence to Jurica Zucko, jzucko@pbf.hr.
* Present address: Hrvoje Petković, AGL Grupo, Instituto de Biomedicina y Biotecnología de Cantabria, Facultad de Medicina, Universidad de Cantabria, Santander, Spain.
Supplemental material for this article may be found at http://aem.asm.org/.

doi:10.1128/AEM.01891-12

the corresponding natural products thus remain out of reach for explorations of their biotechnological and pharmaceutical potentials.

With the development of affordable high-throughput sequencing, it has become attractive to screen genome sequences using bioinformatics tools, such as ClustScan (26), to identify biosynthesis clusters for natural products, which are likely to produce novel chemical entities (an approach often referred to as "genome mining"). The cheapest and simplest approach is to use shotgun sequencing, in which random fragments of genomic DNA are sequenced and the reads are assembled into contigs by programs that look for overlaps. The major problem is that this approach results in a large number of contigs (typically many hundreds). Biosynthesis clusters will typically be split between several contigs, and as there are often several clusters in a single organism, there is no indication of which contigs belong to the same cluster (i.e., are next to each other in the chromosome). A partial answer to this problem is to use paired-end reads, which is significantly more expensive than simple shotgun sequencing. This requires the generation of a set of genomic DNA fragments of almost equal lengths and the sequencing of both ends of each fragment. If the two ends lie in two different contigs, the contigs must be neighbors, and the relative orientation of the two contigs and the approximate length of the nonsequenced gap between them are known. Using paired-end reads, the genome can be assembled into a small number of scaffolds consisting of a series of contigs with known orientation and order. This will usually allow the contigs belonging to one cluster to be identified, but most clusters will be spread over several contigs; i.e., there will be gaps in the sequences of the clusters. The sequences of the gaps can be determined (e.g., by using PCR). Traditional approaches for predicting the chemical structures of secondary metabolites from the DNA sequence require good-quality assembled DNA from the clusters.

The genome sequencing of *Streptomyces* and related genera is significantly more difficult than for most other bacteria. These organisms contain DNA with a high G+C content (over 70%) (12), which often results in shorter reads and a much higher error rate than encountered for the genome sequencing of organisms whose DNA has a lower G+C content. In addition, the high G+C content makes the accidental matching of sequences more likely so that the assembly of genomes from the reads is more difficult. This is exacerbated by the fact that *Streptomyces* genomes are large (7 to 10 Mb) (12) compared to those of most other bacteria. Sequencing errors often result in apparent frameshifting, which causes particular problems for the analysis of large protein-coding genes. Modular biosynthetic enzymes are often encoded by genes that are over 20 kb and are the largest known genes in bacteria. In addition, the multimodular nature of the clusters results in multiple copies of enzymatic domains with extremely similar nucleotide sequences, which can lead to errors during the assembly of the relatively short sequence reads obtained by the high-throughput sequencing techniques into larger contigs.

Streptomyces tsukubaensis NRRL18488 is a natural producer of the medically important immunosuppressant FK506, the biosynthesis of which is encoded by a large biosynthetic cluster containing three modular PKS genes (9, 20). *S. tsukubaensis* NRRL18488 is also the wild-type progenitor of most industrially used FK506-producing strains (16). In this paper, we present an analysis of the PKS and NRPS modular clusters in the genome sequence of *S. tsukubaensis* NRRL18488, which was obtained from simple shotgun reads by using the Roche 454 GS-FLX system (7). We used a phylogenetic approach to determine which contigs belonged to each cluster. The analysis depended heavily on the ClustScan program (26), which allows the semiautomatic analysis of modular biosynthetic clusters with functions that allow the recognition and correction of sequencing errors. Furthermore, with the help of PCR and traditional Sanger sequencing, it was possible to confirm the correct assembly of the sequence contigs. The expression of one of the newly identified clusters was also confirmed by using reverse transcription (RT)-PCR.

MATERIALS AND METHODS

DNA sequencing and assembly. The genomic DNA of *S. tsukubaensis* strain NRRL18488 was sequenced by using the Roche 454 GS-FLX system (7); there were 751,410 reads with an average length of 381 bp. The genome sequence was assembled by using the Newbler package, which is designed specifically for the 454 GS series of pyrosequencing platforms sold by 454 Life Sciences, a Roche Diagnostics company (18), as well as the MIRA 3 whole-genome sequence assembler, which can be used for different sequencing platforms, including Sanger, 454, and Solexa/Illumina (5).

Phylogenetic analysis. The contigs of the assembled genome were translated *in silico* in all six reading frames by using Transeq (http://www.ebi.ac.uk/Tools/emboss/transeq/). The HMMER 2.3.2 program package (8) was used to scan the protein sequences using specially constructed HMM profiles (26) for ketosynthase (KS) domains (PKS clusters) and condensation (C) domains (NRPS clusters) for hits with an expected value of less than 10^{-10}. The protein sequences of hits were extracted from translated contigs by using proprietary Perl and Python scripts. The protein sequences of 372 KS domains (from 36 PKS gene clusters) (see Table S1 in the supplemental material) and 145 C domains (from 21 NRPS gene clusters) (see Table S2 in the supplemental material) were extracted from the CSDB database (http://bioserv.pbf.hr/cms/). Phylogenetic analyses were performed by using the MEGA 5 software package (27). Two tree construction methods were used: neighbor joining and minimal evolution. Distances were computed by using the JTT matrix and the pairwise deletion option. The minimal evolution method started with a neighbor-joining tree and used the close-neighbor-interchange (CNI) algorithm at a search level of zero. Bootstrap analyses with 1,000 replicates were performed.

Annotation of gene clusters and prediction of products. The ClustScan program package (26) was used to annotate the clusters. For PKS clusters, ClustScan predicts the chemical structure of the linear backbone. For NRPS clusters, the adenylation (A)-domain amino acid sequences were extracted and compared to a set of 397 A domains of known specificity by using a latent semantic indexing method (D. Baranasic et al., unpublished data) to identify the most similar domain, which was used to predict the incorporated amino acid.

DNA isolation, manipulation, and cloning. Culture conditions for *S. tsukubaensis* NRRL18488 were described previously (9). The isolation and manipulation of genomic DNA were carried out according to standard methods (15, 23). PCRs were carried out by using AccuPrime GC-rich DNA polymerase (Invitrogen), according to the standard protocol provided by the manufacturer. Primer sequences and specific annealing temperatures are shown in Table S3 in the supplemental material. PCR products were cloned by using the TA cloning kit (Invitrogen) and sequenced by Macrogen, Inc. Most of the PCR products did not give good results with the standard sequencing service, so they were sequenced with a proprietary "difficult sequencing protocol" (Macrogen, Inc.), using modified deoxynucleotide triphosphates (6).

RNA isolation and purification. As described previously, a seed culture was used for the inoculation of 250-ml Erlenmeyer flasks containing 50 ml of PG3 production medium (9), and cultivation was carried out at 28°C at 220 rpm for 62 h. A 2-ml sample was collected and mixed thoroughly with 4 ml of RNAlater RNA stabilization reagent (Qiagen). After 5 min of incubation at room temperature, the mixture was stored at −80°C.

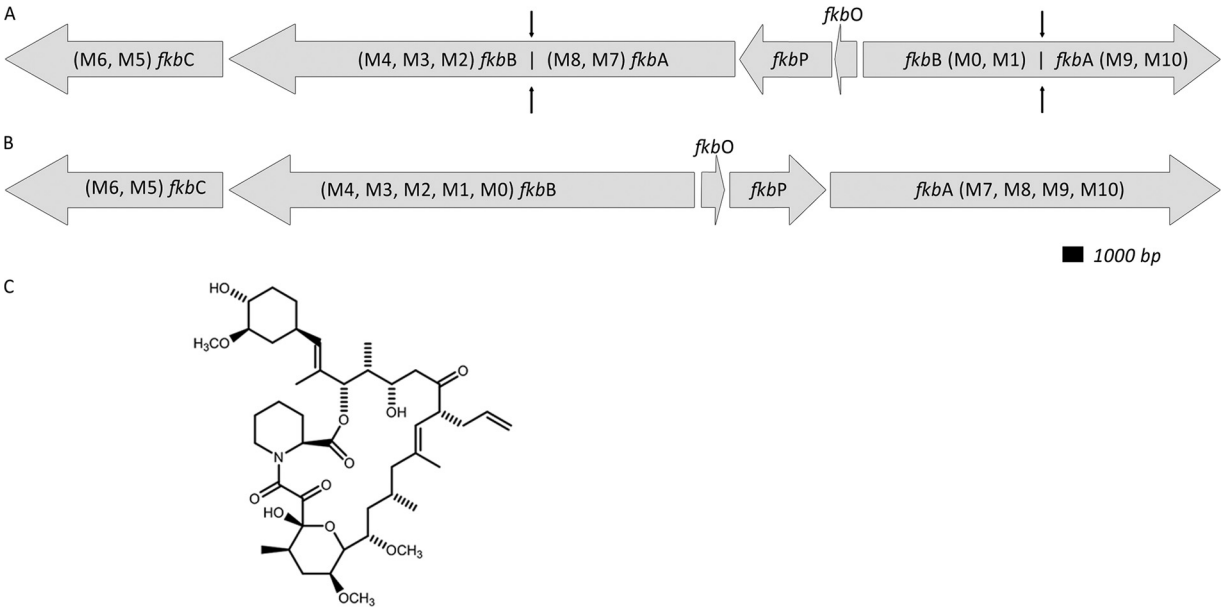

FIG 1 Apparent inversion in the FK506 cluster. (A) Module organization as deduced by the ClustScan program using module numbering compatible with the annotation of the FK520 cluster. (B) Module organization after correction of the apparent inversion introduced by an assembly error. (C) Chemical structure of FK506.

Cell pellets were thawed, centrifuged, and homogenized with FastPrep Lysing Matrix B tubes and a FastPrep-24 instrument (MPBio). Total RNA was extracted by using the RNeasy Midi kit (Qiagen). Isolated samples were treated with DNase I (Invitrogen), according to the manufacturer's instructions. The integrity of total RNA was assessed by agarose gel electrophoresis to visualize 16S and 23S rRNA genes, and the purity of total RNA was assessed spectrophotometrically by the A_{260}/A_{280} ratio.

RT-PCR analysis. Reverse transcription was carried out with Maxima reverse transcriptase (Fermentas), according to the manufacturer's protocol, using random hexamer primers and 5 μg of total RNA as the template. PCRs were carried out on the cDNA using pairs of primers specific for the PKS2 and PKS4 genes, respectively (see Table S3 in the supplemental material) and Phusion high-fidelity DNA polymerase (Finnzymes). Thirty-two cycles of DNA amplification were used in all cases. The *hrdB* gene, which encodes the major sigma factor, was used as a positive control, as it is thought to be expressed at a constant level (4). As a negative control, RNA samples which had not been treated with reverse transcriptase were used as the templates for PCR, to confirm that the amplified products were not derived from chromosomal DNA. The PCR products were analyzed by electrophoresis on a 1.2% agarose gel. Two biological replicates were used, each with two experimental replicates to confirm reproducibility. The primers gave single bands of the expected size with genomic DNA (data not shown).

Nucleotide sequence accession numbers. The annotated sequences were deposited in GenBank as 24 entries under accession no. JX081645 to JX081668.

RESULTS

Identification of PKS and NRPS clusters. Total DNA of *Streptomyces tsukubaensis* NRRL18488 was used for shotgun sequencing with a 36-fold genome coverage. The genome was initially assembled by using the Newbler program (18). However, a preliminary analysis showed that the FK506 cluster was distributed over five contigs. A second assembly was undertaken by using the MIRA 3 program (5), which produced a single contig containing the whole FK506 cluster. For further analyses, it was decided that both assemblies should be used, because although the MIRA 3 program performed better for the FK506 cluster, the Newbler assembler produced fewer contigs (240 compared to 912) and larger contigs (N_{50} parameter of 77,592 bp compared to 46,295 bp; i.e., 50% of base pairs are in contigs larger than the N_{50} value) for the whole genome. In order to identify potential modular PKS and NRPS clusters, the sequences were translated in all six reading frames, and the most highly conserved domains of PKS and NRPS modules (the KS and C domains, respectively) (24) were detected by using programs of the HMMER package (8) with KS and C domain profiles constructed for the ClustScan program (26). A nonstringent cutoff value was used (expected value of less than 10^{-10}) to allow the detection of interrupted sequences at the ends of contigs and unusual domains. This analysis detected 60 putative KS domains and 46 putative C domains. Contigs containing potential domains were viewed by using the ClustScan program, and putative modular clusters were identified by the presence of several domains. Most of the putative C domains (41/46 domains) and about two-thirds of the putative KS domains (38/60 or 41/61 domains, depending on the assembly) could be assigned to such clusters. The remaining ones had low scores due to poor matching or only short pieces of the domains being present (e.g., they were on singleton reads) and were not analyzed further.

The FK506 cluster was identified by using its similarity to the FK520 cluster (29), because the DNA sequence of a cluster producing FK506 (19) had not yet been published. In the MIRA 3 assembly, it was present on a single contig (c216). The cluster was analyzed by using the ClustScan program but surprisingly did not have a structure compatible with the production of FK506 (Fig. 1A). In particular, one gene carried the loading domain (module 0) and module 1, followed by the last two modules of the cluster (modules 9 and 10). Another gene carried modules 7 and 8, followed by modules 2, 3, and 4. A closer examination showed that the inversion of a 15.8-kb region in the middle of the sequence would generate a cluster compatible with the production of FK506

TABLE 1 PKS and NRPS modular gene clusters present in the genome of *S. tsukubaensis* NRRL18488 showing the distribution between the Newbler and MIRA 3 contigs

Cluster	Newbler		MIRA 3	
	No. of contigs	No. of modules	No. of contigs	No. of modules
PKS genes				
FK506	5	10	1	9
St-PKS2	2	11	3	12
St-PKS3	10	12	10	15
St-PKS4	3	5	3	5
NRPS genes				
St-NRPS1	1	21	1	21
St-NRPS2	1	6	6	6
St-NRPS3	3	3	3	3
St-NRPS4	3	6	3	6
St-NRPS5	2	3	2	3
St-NRPS6	3	5	5	5

(Fig. 1B). The sequences at the ends of the inversion form a nearly perfect (99% nucleotide identity) inverted repeat of nearly 1 kb. The apparent inversion seems to be an assembly artifact, because an assembler program cannot determine the correct orientation if the length of the inversion is longer than the read length. It is also not possible to deduce the exact sequences of the two repeats, i.e., whether they are totally identical over a considerable length. The assembly of the DNA sequence was corrected by removing the inversion and showed 99% identity to the recently reported sequence of the FK506 cluster (GenBank accession no. HM116537) from *Streptomyces* sp. strain KCTC 11604 BP (19).

Because modular clusters are large, the domains of a single biosynthesis cluster are often distributed between more than one contig in an assembly. The KS domains (PKS clusters) were distributed over 16 to 19 contigs, and the C domains (NRPS clusters) were distributed over 13 to 20 contigs, depending on the assembly used. An effective analysis of the biosynthetic clusters depends on the ability to identify which contigs carry modules of the same cluster. As only simple shotgun sequencing data were available, there were no experimental data to indicate which contigs were next to each other in the chromosome. We developed a bioinformatics method based on phylogenetic trees to identify contigs carrying modules from the same cluster, thus obviating the need to obtain further laboratory data. When phylogenetic trees of KS or C domains were examined, it was found that the domains from a particular cluster usually group together (14), which is probably due to gene conversion events (31). The KS and C domains identified in the translated *S. tsukubaensis* genome sequence were used to construct phylogenetic trees together with 372 KS and 145 C domains from well-characterized gene clusters (36 PKS and 21 NRPS clusters from CSDB [http://bioserv.pbf.hr/cms/]) (see Fig. S1 and S2 in the supplemental material). In most cases, domains that were clustered on a single contig were also clustered in the phylogenetic tree. When more than one contig contained domains which clustered together in the phylogenetic tree, they were assigned to the same cluster. These analyses resulted in the prediction of 4 PKS and 6 NRPS clusters (Table 1). Two different methods of tree construction produced the same conclusion. Although a small number of modules may not group with others of their cluster in the phylogenetic trees, they are rare enough to pose little problem in assembling the clusters.

Table 1 shows that the two assemblies used sometimes differed in the degree of fragmentation of clusters between contigs: the MIRA 3 assembly contained the entire FK506 cluster in a single contig, whereas the Newbler assembly split the cluster between five contigs. In contrast, the *S. tsukubaensis* NRPS2 (*St*-NRPS2) cluster was present on a single contig in the Newbler assembly but was scattered over six contigs in the MIRA 3 assembly. With the exception of the *St*-PKS3 cluster, which was distributed over 10 contigs in each assembly, every cluster could be found in maximally three contigs by using the appropriate assembly of the genome.

Characterization of clusters. The PKS gene cluster *St*-PKS2 was present on a single contig in the Newbler assembly. Analysis with ClustScan suggested that there were five genes present, and domains could be assembled into modules in a typical pattern for a modular PKS cluster. The overview presented by ClustScan allows a critical assessment of the quality of the sequencing and assembly. Mistakes can result in apparent frameshifts, which would be interpreted as the splitting of genes into smaller apparent genes (26). However, in this case, the gene boundaries coincide with module boundaries, indicating that the assignment to genes is correct. ClustScan predicts the substrate specificity of AT domains and the activity and stereochemical specificity of KR domains. It also predicts whether DH and ER domains are functional. This degree of annotation helps in the comparison of clusters. In the case of *St*-PKS2, the phylogenetic analysis indicated similarity to concanamycin A (see Fig. S1 in the supplemental material). However, *St*-PKS2 contained only 11 extension modules, whereas the concanamycin A cluster had 13 extension modules. ClustScan (26) was used to annotate both clusters in a uniform way, and the predicted specificities were used to compare the two clusters. In the case of concanamycin A, ClustScan predicted that the AT domains of modules 4 and 6 would use an ethylmalonyl coenzyme A (CoA) substrate, whereas the chemistry of the product suggests that methylmalonyl-CoA and methoxymalonyl-CoA substrates are actually used (11). A comparison of the two clusters showed that eight module pairs (modules 1 to 8 of *St*-PKS2 and modules 2 to 9 of the concanamycin A cluster) had identical predicted specificities for the AT and KR domains, except that module 3 of *St*-PKS2 lacked a KR domain (a detailed analysis of the *St*-PKS2 cluster is shown in Fig. S3 in the supplemental material). Moreover, they are distributed in the same way over three genes. The only other differences between the corresponding modules are that modules 4 and 5 of *St*-PKS2 lacked a DH domain. The other modules of the two clusters (modules 9 to 11 of *St*-PKS2 and modules 10 to 13 of concanamycin A) did not show any such similarity in specificities.

Analysis of the PKS cluster *St*-PKS4 presents a more difficult challenge, as it is distributed over 3 contigs in each assembly. In an initial step, all PKS domains in the three MIRA 3 contigs were detected by using ClustScan (Fig. 2). It was noticed that domains were distributed throughout the small contig c213, whereas they were present only near a single end in both contigs c43 and c288, suggesting that contig c213 should be located in the middle between the other two contigs. This putative assembly of the three contigs was supported by a detailed examination of the domains, which are encoded on the lower strand (Fig. 2) in all three contigs. At the end of contig c288, we identified a KS domain and an AT

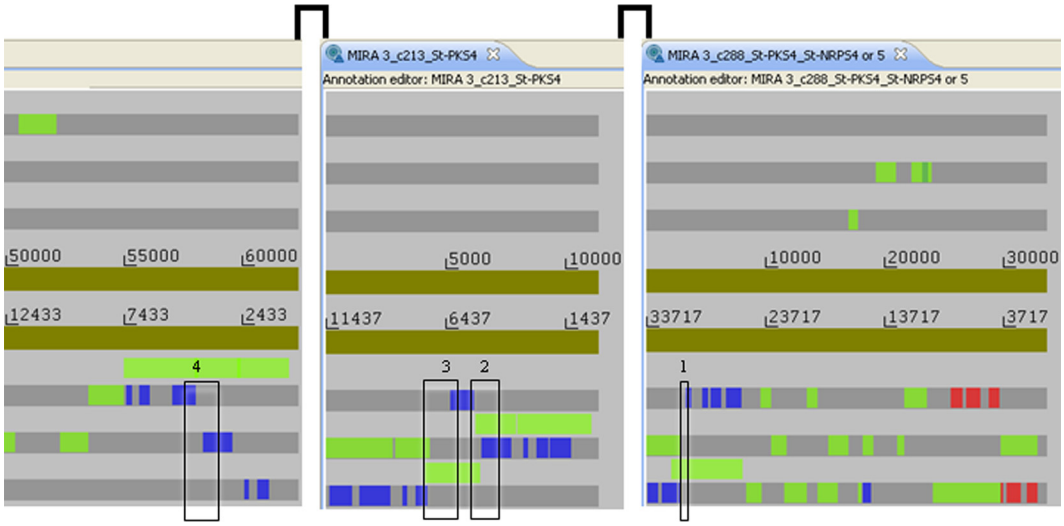

FIG 2 Screen shots of the three MIRA 3 contigs c43, c213, and c288, which contain cluster *St*-PKS4. The modules in the contigs were identified as belonging to the same cluster by phylogenetic analysis. The positions of genes and domains are shown on the three reading frames of the forward strand (top part of the windows) and the three reading frames of the complementary strand (bottom part of the windows). Thus, in this case, most of the domains are on the complementary strand, distributed over all three reading frames. The PKS domains are shown as blue boxes. By moving the cursor over the boxes, it is possible to see what sort of domain has been detected and whether it is complete. As the order of the domains in modules is highly conserved, it is easy to see whether the sequence in another contig is a candidate for the continuation of a cluster; gaps between contigs are usually small compared to the module size. The postulated order and orientation of the three contigs are indicated. The four numbered boxes contain the apparent frameshifts that cause the reading frames to change within modules. For clarity, the left end of c43 and the right end of c288 are omitted. The green and red boxes show predicted proteins and NRPS domains, respectively.

domain, while the ER, KR, and ACP domains were identified at the beginning of contig c213. This suggests the existence of a module spanning the gap between the contigs. Similarly, there appears to be a module spanning the gap between contigs c213 and c43. The assumption that the three contigs represent parts of the same cluster was tested by designing PCR primers to amplify the sequences in the predicted gaps between the contigs. In both cases, a PCR product (1,459 bp and 1,695 bp, respectively) could be cloned and sequenced (data not shown) and proved that the contigs are joined, as suggested in the legend of Fig. 2. It was possible to obtain PCR products only by using a special DNA polymerase optimized for GC-rich sequences, and the sequencing of the products was successful only when a protocol for difficult DNA sequences was used (see Materials and Methods).

The KS, AT, KR, and ACP domains are present in the middle of contig c213, which suggests the presence of a further module (Fig. 2). However, these domains are distributed over three different reading frames. There are also two further apparent frameshifts in contigs c213 and c288, respectively (Fig. 2). PCR primers were designed to amplify the regions around the four apparent frameshifts. The PCR products were cloned and sequenced, and it was confirmed in all four cases that the apparent frameshifts were not present; i.e., they were due to sequencing or assembly errors. After the linking of the contigs and correction of these errors, it was possible to perform a detailed analysis of the structure of cluster *St*-PKS4 by using ClustScan.

There were four apparent frameshifts in cluster *St*-PKS4. However, the FK506 cluster and cluster *St*-PKS2 did not contain any obvious frameshifts. The contigs of the Newbler assembly were translated and searched with 830 HMM profiles constructed for common bacterial genes, in order to assess the frequency of sequencing errors resulting in apparent frameshifts. Good hits were obtained with 555 HMM profiles. However, for 71 of these hits (13%), the alignments to the profile were distributed between different reading frames, showing apparent frameshifts (data not shown). As it is expected that nearly all of these genes will be intact in a wild-type organism, these apparent frameshifts should be due to sequencing or assembly artifacts.

All six of the NRPS clusters were distributed among 1 to 3 contigs for at least one of the assemblies used, so they could be assembled and analyzed further. The NRPS modular gene cluster *St*-NRPS1 was present on a single contig in both assemblies, and the Newbler contig was analyzed further. The *St*-NRPS1 gene cluster has three genes, which are probably organized in the biosynthetic order, because the first gene has a loading module at the beginning and the third gene has a TE domain at the end. There are 18 extender modules (Table 1). The C domains of the cluster were grouped close to those of enduracidin (30), which has 16 extender modules, in the phylogenetic tree. It is interesting that both clusters contain one module that lacks an A domain, module 7 of enduracidin and module 5 of *St*-NRPS1. The module architectures of the assembled PKS and NRPS clusters are shown in Fig. S5 in the supplemental material.

Prediction of chemical products and expression. ClustScan can predict the chemical structures of the products of modular PKS clusters. After the assembly of the three contigs and correction of the frameshifts, we predicted the structure of the linear backbone of the product of cluster *St*-PKS4 (Fig. 3A). Although the KS domains of *St*-PKS4 were close to those of nigericin (10) in the phylogenetic trees, the structures of the two compounds do not show any obvious similarity. The putative product of cluster *St*-PKS2 was also predicted (Fig. 3B). This would be expected to form a macrocyclic ring analogous to that of concanamycin A (11), and the resulting ring structure would be similar to that of bafilomycin (28) (see Fig. S4 in the supplemental material); several different bafilomycins are known to be produced by various

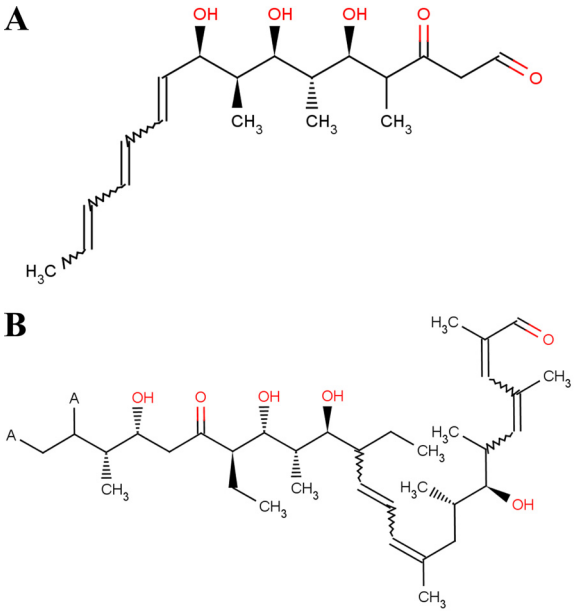

FIG 3 Predicted chemical structures of linear products of modular clusters. (A) Cluster *St*-PKS4. (B) Cluster *St*-PKS2. This structure should cyclize to form a bafilomycin. The generic side chains represented by A indicate parts of the molecule for which ClustScan cannot make a specificity predication. (C) Cluster *St*-NRPS1. ??? corresponds to a module for which no substrate prediction was possible.

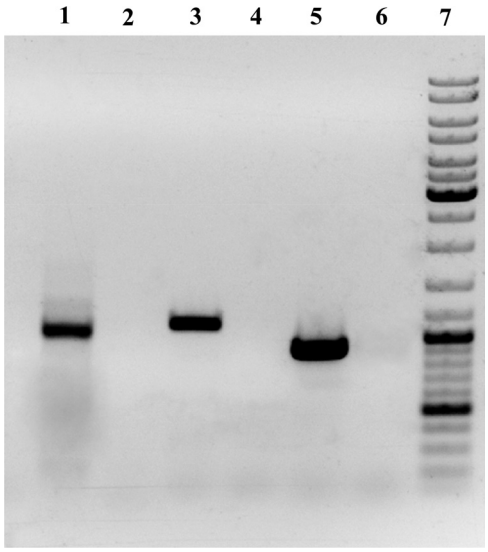

FIG 4 Transcription of cluster *St*-PKS2. Lanes 1, 3, and 5, RT-PCR with primers for the *pks2*, *pks4*, and *hrdB* genes; tracks 2, 4, and 6, controls without reverse transcriptase; track 7, GeneRule 100- to 10,000-bp DNA ladder (Fermentas).

Streptomyces species. It was predicted that methoxymalonate would be incorporated by modules 5 and 11 of *St*-PKS2 (Fig. 3B). Unlike the common extender substrates (malonyl-CoA and methylmalonyl-CoA), which are probably available in all *Streptomyces* species, the supply of methoxymalonate requires a special pathway. Homologues of genes involved in the supply of methoxymalonate were identified immediately downstream of the last PKS gene of cluster *St*-PKS2. Module 3 of *St*-PKS2 was predicted to incorporate ethylmalonyl-CoA, but no genes involved in the supply of this substrate could be identified in the cluster. However, we have previously shown that the ethylmalonyl-CoA extender unit is readily available in *S. tsukubaensis* cells (16).

The chemical structure of NRPSs is determined mainly by the specificity of the A domains. Recently, we implemented a prediction program for A-domain specificity based on latent semantic indexing (Baranasic et al., unpublished). This was used to suggest a chemical structure for the product of *St*-NRPS1 (Fig. 3C). The absence of an A domain in the sixth module prevents predictions of the complete peptide linear structure.

Although the presence of intact biosynthetic clusters suggests that they should synthesize the corresponding products, it is also possible that they are silent clusters. We used RT-PCR to examine whether the bafilomycin-like cluster (*St*-PKS2) was transcribed. There are five PKS genes, which are all carried on the same DNA strand, and primers were designed to detect transcripts of the second and fourth PKS genes. Previous experience using fermentations of the strain to produce FK506 (16) suggested that 62 h of fermentation would be a suitable time to detect the expression of secondary metabolite gene clusters. Figure 4 shows that both genes are transcribed under the growth conditions that were used for this experiment.

DISCUSSION

The generation of high-quality DNA sequences is disproportionately time-consuming and expensive. It is therefore important to develop methods for the analysis of lower-quality sequences. In the case of modular biosynthetic clusters, there are two major problems: as the clusters are large, they will often be distributed over multiple contigs, and sequencing errors resulting in apparent frameshifts will disturb the analysis of the large proteins characteristic for modular PKSs and NRPSs. In this paper, we characterized clusters in *S. tsukubaensis* NRRL18488 using simple read data (i.e., no paired-end reads were used). Two different assembly programs were used, because some clusters were assembled better with one program than the other (Table 1): Newbler (18) performed better on the *St*-PKS2 cluster, whereas MIRA 3 (5) performed better on the FK506 gene cluster. As the modules have a degree of sequence similarity to each other, it seems likely that the presence of imperfect repeated sequences might have affected the performance of the assembly programs. In the case of FK506, the presence of a near-perfect inverted repeat of nearly 1 kb resulted in an apparent inversion in the cluster (Fig. 1). Such a situation cannot be resolved by an assembly program but was easy to detect by using the ClustScan program, which gives a visual representation of the module and domain structure of the clusters (26).

Clusters could be easily identified by using HMM profile searches for the most conserved domains of PKS (KS domains) and NRPS (C domains). A method was developed to assign contigs to clusters based on the clustering of domains from one cluster in a phylogenetic tree. In 9/10 cases, the clusters were present on maximally three contigs. In such cases, the analysis of the module

structure with ClustScan suggests the order and respective orientation of the contigs (Fig. 2). The analysis also allows an easy visual identification of apparent frameshifts. This allows the design of PCR primers so that the sequencing of PCR products can close gaps between contigs and correct frameshifts. Even in the absence of an experimental verification of the cluster structure, ClustScan can predict most of the chemistry of the putative product, so it can be decided whether the product is interesting enough to justify further experimental work.

A recent paper presented the genome sequence of *S. tsukubaensis* NRRL18488 (2). A combination of shotgun sequencing, paired-end reads, and comparisons with a cosmid gene bank allowed the 959 contigs to be ordered in the chromosome. Ten modular clusters were found, which is in agreement with our analyses (Table 1). However, three clusters were annotated as PKS-NRPS hybrid clusters, whereas we had classified them as NRPS clusters. This is explained by the fact that we used the NRPS A domains to identify the cluster modules and had not carried out a detailed analysis of the clusters. A traditional pipeline was used for annotation. Because nearly all the clusters had gaps in their sequences (i.e., they were distributed over several contigs), the antiSMASH annotation program (17) could not be used to predict the chemical structures of the products. It is striking that a combination of the phylogenetic method to identify contigs belonging to a single cluster as well as the use of the visualization offered by ClustScan to assemble clusters and detect sequencing errors allowed us to extract most of the relevant information about modular clusters from simple and affordable shotgun sequencing. Once clusters have been assembled and corrected by using ClustScan, it is possible to use any other available analysis program for extending the analysis (e.g., antiSMASH [17]).

Four sequencing errors that produced apparent frameshifts were identified in cluster *St*-PKS4 (Fig. 2). However, clusters FK506 and *St*-PKS2 did not show any frameshifts. An analysis of 555 genes that are present in many bacteria showed that 13% contained apparent frameshifts (data not shown), probably due to sequencing and assembly errors. This would correspond to about one such error in ca. 10 kb. The DNA for genome sequencing was prepared from exponentially growing cells, so sequences near the origin of replication should be present at a higher copy number than those near the chromosome ends (21). This would allow a better assembly of sequences closer to the origin.

The prediction of chemical structures based on DNA sequences will always include a degree of error and uncertainty but can give sufficient information to indicate whether a cluster is of potential interest (26). We developed a method for predicting the specificity of A domains, which is the major factor determining the product chemistry in NRPSs, and used it to predict the product of cluster *St*-NRPS1 (Fig. 3C). We predicted the products of two PKS clusters using ClustScan (Fig. 3A and B), and this showed that *St*-PKS2 probably produces a structure similar to that of bafilomycin. A comparison of the cluster architecture with that of concanamycin A showed that some pairs of modules (modules 1 to 8 of *St*-PKS2 and modules 2 to 9 of concanamycin A) were very similar to each other and were distributed in an identical fashion among three genes. However, the other modules did not show such similarity, suggesting that the two clusters might have diverged through recombination with another unrelated cluster. Such single-crossover events were suggested previously to be a major driving force in the evolution of clusters (24, 25, 31). It was possible to demonstrate that cluster *St*-PKS2 was transcribed by using RT-PCR (Fig. 4). This approach offers a general strategy to isolate novel metabolites. Bioinformatics would be used to predict the chemical structures of products, allowing the identification of promising target clusters. RT-PCR would define fermentation conditions under which such clusters are transcribed. The purification of the novel products from fermentation broths would also be assisted by the chemical information available.

This work exemplifies how relatively low-cost sequences of *Streptomyces* genomes obtained by high-throughput techniques can lead to the efficient and rapid assembly of sequences of most complex modular biosynthetic genes by using appropriate bioinformatics tools. The results of such analyses can greatly facilitate the identification of novel compounds by genome mining as well as lead to a rapid improvement of the yield of target compounds by using metabolic engineering or classical strain improvement approaches. The ClustScan program offers particular advantages, as it is designed to be semiautomatic, allowing extensive user intervention, which is necessary to obtain useful data despite sequencing errors and problems with assemblies.

ACKNOWLEDGMENTS

We thank the Government of Slovenia, Ministry of Higher Education, Science and Technology (Slovenian Research Agency [ARRS]), for the award of grant no. J4-9331 and L4-2188 to H.P. We also thank the Ministry of the Economy, the JAPTI Agency, and the European Social Fund (contract no. 102/2008) for the funds awarded for the employment of G.K. This work was also funded by a cooperation grant of the German Academic Exchange Service (DAAD) and the Ministry of Science, Education, and Sports, Republic of Croatia (to J.C. and D.H.), and by grant 09/5 (to D.H.) from the Croatian Science Foundation.

REFERENCES

1. **Baltz RH.** 2006. Marcel Faber Roundtable: is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? J. Ind. Microbiol. Biotechnol. **33**:507–513.
2. **Barreiro C, et al.** 2012. Draft genome of *Streptomyces tsukubaensis* NRRL 18488, the producer of the clinically important immunosuppressant tacrolimus (FK506). J. Bacteriol. **194**:3756–3757.
3. **Bentley SD, et al.** 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature **417**:141–147.
4. **Buttner MJ, Chater KF, Bibb MJ.** 1990. Cloning, disruption, and transcriptional analysis of three RNA polymerase sigma factor genes of *Streptomyces coelicolor* A3(2). J. Bacteriol. **172**:3367–3378.
5. **Chevreux B, et al.** 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. **14**:1147–1159.
6. **Dierick H, Stul M, De Kelver W, Marynen P, Cassiman J-J.** 1993. Incorporation of dITP or 7-deaza dGTP during PCR improves sequencing of the product. Nucleic Acids Res. **21**:4427–4428.
7. **Droege M, Hill B.** 2008. The Genome Sequencer FLX system—longer reads, more applications, straight forward bioinformatics and more complete data sets. J. Biotechnol. **136**:3–10.
8. **Eddy SR.** 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. **23**:205–211.
9. **Goranovič D, et al.** 2010. Origin of the allyl group in FK506 biosynthesis. J. Biol. Chem. **285**:14292–14300.
10. **Harvey BM, et al.** 2007. Insights into polyether biosynthesis from analysis of the nigericin biosynthetic gene cluster in *Streptomyces* sp. DSM4137. Chem. Biol. **14**:703–714.
11. **Haydock SF, et al.** 2005. Organization of the biosynthetic gene cluster for the macrolide concanamycin A in *Streptomyces neyagawaensis* ATCC 27449. Microbiology **151**:3161–3169.
12. **Hopwood DA.** 2006. Soil to genomics: the *Streptomyces* chromosome. Annu. Rev. Genet. **40**:1–23.
13. **Ikeda H, et al.** 2003. Complete genome sequence and comparative anal-

ysis of the industrial microorganism *Streptomyces avermitilis*. Nat. Biotechnol. **21**:526–531.
14. **Jenke-Kodama H, Börner T, Dittmann E.** 2006. Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. PLoS Comput. Biol. **2**:e132. doi:10.1371/journal.pcbi.0020132.
15. **Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA.** 2000. Practical *Streptomyces* genetics. The John Innes Foundation, Norwich, United Kingdom.
16. **Kosec G, et al.** 2012. Novel chemobiosynthetic approach for exclusive production of FK506. Metab. Eng. **14**:39–46.
17. **Medema MH, et al.** 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. **39**:W339–W346. doi: 10.1093/nar/gkr466.
18. **Miller JR, Koren S, Sutton G.** 2010. Assembly algorithms for next-generation sequencing data. Genomics **95**:315–327.
19. **Mo S, et al.** 2011. Biosynthesis of the allylmalonyl-CoA extender unit for the FK506 polyketide synthase proceeds through a dedicated polyketide synthase and facilitates the mutasynthesis of analogues. J. Am. Chem. Soc. **133**:976–985.
20. **Motamedi H, Shafiee A.** 1998. The biosynthetic gene cluster for the macrolactone ring of the immunosuppressant FK506. Eur. J. Biochem. **256**:528–534.
21. **Musialowski MS, et al.** 1994. Functional evidence that the principal DNA replication origin of the *Streptomyces coelicolor* chromosome is close to the *dnaA-gyrB* region. J. Bacteriol. **176**:5123–5125.
22. **Nett M, Ikeda H, Moore BS.** 2009. Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat. Prod. Rep. **26**:1362–1384.
23. **Sambrook J, Russell DW.** 2001. Molecular cloning: a laboratory manual, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
24. **Starcevic A, et al.** 2011. A novel docking domain interface model that can predict recombination between homoeologous modular biosynthetic gene clusters. J. Ind. Microbiol. Biotechnol. **38**:1295–1304.
25. **Starcevic A, et al.** 2012. Recombinatorial biosynthesis of polyketides. J. Ind. Microbiol. Biotechnol. **39**:503–511.
26. **Starcevic A, et al.** 2008. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. Nucleic Acids Res. **36**:6882–6892.
27. **Tamura K, et al.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28**:2731–2739.
28. **Werner G, Hagenmaier H, Drautz H, Baumgartner A, Zähner H.** 1984. Metabolic products of microorganisms. 224. Bafilomycins, a new group of macrolide antibiotics. Production, isolation, chemical structure and biological activity. J. Antibiot. (Tokyo) **37**:110–117.
29. **Wu K, Chung L, Revill WP, Katz L, Reeves CD.** 2000. The FK520 gene cluster of *Streptomyces hygroscopicus* var. *ascomyceticus* (ATCC 14891) contains genes for biosynthesis of unusual polyketide extender units. Gene **251**:81–90.
30. **Yin X, Zabriskie TM.** 2006. The enduracidin biosynthetic gene cluster from *Streptomyces fungicidicus*. Microbiology **152**:2969–2983.
31. **Zucko J, Long PF, Hranueli D, Cullum J.** 2012. Horizontal gene transfer drives convergent evolution of modular polyketide synthases. J. Ind. Microbiol. Biotechnol. **39**:1541–1547.